DOCUMENT RESUME

ED 064 362 TM 001 530

AUTHOR Koplyay, Janos B.
TITLE Multiple Regression Analysis and Automatic Interaction Detection.
NOTE 10p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Branching; *Correlation; *Mathematical Models; *Multiple Regression Analysis; *Predictor Variables; Statistical Analysis
IDENTIFIERS AID; *Automatic Interaction Detector

ABSTRACT

The Automatic Interaction Detector (AID) is discussed as to its usefulness in multiple regression analysis. The algorithm of AID-4 is a reversal of the model building process; it starts with the ultimate restricted model, namely, the whole group as a unit. By a unique splitting process maximizing the between sum of squares for the categories of each variable while minimizing the error sum of squares (within group sum of squares), AID-4 seeks out that variable which has the largest between sum of squares and splits the original group into two mutually exclusive groups on this variable at that category where the maximum between sum of squares occurred. The major advantage of using AID-4 is that the maximum squared composite correlation is obtained without the task of attempting to identify the various relevant combinations of linear and non-linear interaction terms by trial and error necessary in the full model of the multiple regression technique. (Author/DB)

ED 064362

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

# MULTIPLE REGRESSION ANALYSIS AND AUTOMATIC INTERACTION DETECTION

By

Janos B. Koplyay, Ph.D.
Personnel Research Division
Air Force Human Resources Laboratory (AFSC)
Lackland Air Force Base, Texas

TM 001 530

Multiple regression analysis is a powerful approach to the formulation and the analysis of research problems, and the testing of hypotheses. It is less restrictive than multiple correlational analysis; e.g., multiple regression analysis does not assume that the predictor variables constitute a multivariate normal distribution. The absence of this restriction permits the introduction of categorical predictor variables. One use for such variables is the establishment of mutually exclusive groups and the testing of the hypothesis that knowledge of group membership at different levels of a predictor variable improves the accuracy of prediction of a criterion of interest. The automatic interaction detector improves the power and efficiency of the application of multiple regression analysis through the identification of optimal configurations of predictor variables for criterion prediction. Joint familiarity with regression techniques and the application of the automatic interaction detector will provide the research scientist with an effective tool. Without the automatic interaction detector, the establishment of optimally effective sets of predictor variables is essentially a cut-and-try, guesswork process. With automatic interaction detection, guidance is offered directly as to the optimal prediction possible with the predictor set, and the identification of

reduced subsets of predictors which most closely approximate the total validity of the full set of predictors. In this sense, AID-4 is a model identifying process.

The multiple regression technique as illustrated by Bottenberg and Ward (1963), starts with a K-category full regression model including all the predictor variables (categorical and/or continuous) and the basic procedure consists of testing for the significance of the difference between the error sum of squares resulting when some of the least-square weighted categorical memberships are not taken into account in the (K-n)-category restricted model where n is the number of restrictions imposed upon the full model. The test of significance is done by the F-statistic, comparing the minimized error sum of squares of the full model with that of the restricted model. This comparison indicates the extent to which the eliminated n categorical memberships contributed to the accuracy of predicting the criterion variable.

For a simple example, let us suppose that we have two predictor variables $x^{(1)}$ with three levels, i.e., high school degree, undergraduate degree and graduate degree; and $x^{(2)}$ with two levels, i.e., pilot or navigator. The criterion variable is some test score on a 50-item test and we have 60 individuals in the experiment. (The actual data was taken from an example in Hays' Statistic, Holt, Rinehart and Winston, 1963, p. 403.) The simple two predictor, one criterion multiple regression model is:

TM 001 530

$$\underline{\text{Model 1}} \qquad y = a_0 u + a_1 x^{(1)} + a_2 x^{(2)} + e_1$$

which after the conventional multiple regression yields a solution of $R_1^2 = .7508$ and a mimimized error sum of squares of $q_1 = 1607.4670$.

Testing for interaction one would include a product term in the model:

$$\underline{\text{Model 2}} \qquad y = b_0u + b_1x^{(1)} + b_2x^{(2)} + b_3x^{(1)} \cdot x^{(2)} + e_2$$

Model 2 is the so called "full model" and Model 1 is the "restricted model." It is restricted because we impose the restriction of $b_3 = 0$ upon Model 2 thus obtaining Model 1. By comparing the minimized error sums of squares of Model 1 and Model 2, $q_1$ and $q_2$ respectively, one gets an indication of the contribution of the product term (or "interaction") to the predictive efficiency of the system. The solution of Model 2 gives an $R_2^2 = .8184$ and $q_2 = 1171.8683$.

The F-statistic is computed by:

$$F = \frac{(q_1 - q_2)/(4 - 3)}{q_2/(60 - 4)} = 20.82$$

with df = 1 and 56. We can make further "guesses" about the predictor variables. Let us assume that predictor $x^{(1)}$ has a quadratic component and that the previously hypothesized interaction is also present. Our model will look like:

$$\underline{\text{Model 3}} \qquad y = c_0u + c_1x^{(1)} + c_2x^{(2)} + c_3x^{(1)} \cdot x^{(2)} + c_4 \cdot \left[x^{(1)}\right]^2 + e_3$$

The solution of Model 3 yields an $R_3^2 = .8423$ and a minimized error sum of squares $q_3 = 1017.7627$. The F-statistic is:

$$F = \frac{(q_2 - q_3)/(5 - 4)}{q_3/(60 - 5)} = 8.33$$

with df = 1 and 55. Additional possible models are listed below:

Model 4 $y_4 = d_0u + d_1x^{(1)} + d_2x^{(2)} + d_3x^{(1)}x^{(2)} + d_4\left[x^{(2)}\right]^2 + e_4$

$R_4^2 = .8184$ $q_4 = 1171.8683$

Model 5 $y_5 = k_0u + k_1x^{(1)} + k_2x^{(2)} + k_3x^{(1)}x^{(2)} + k_4\left[x^{(1)}\right]^2 + k_5\left[x^{(2)}\right]^2 + e_5$

$R_5^2 = .8423$ $q_5 = 1017.7627$

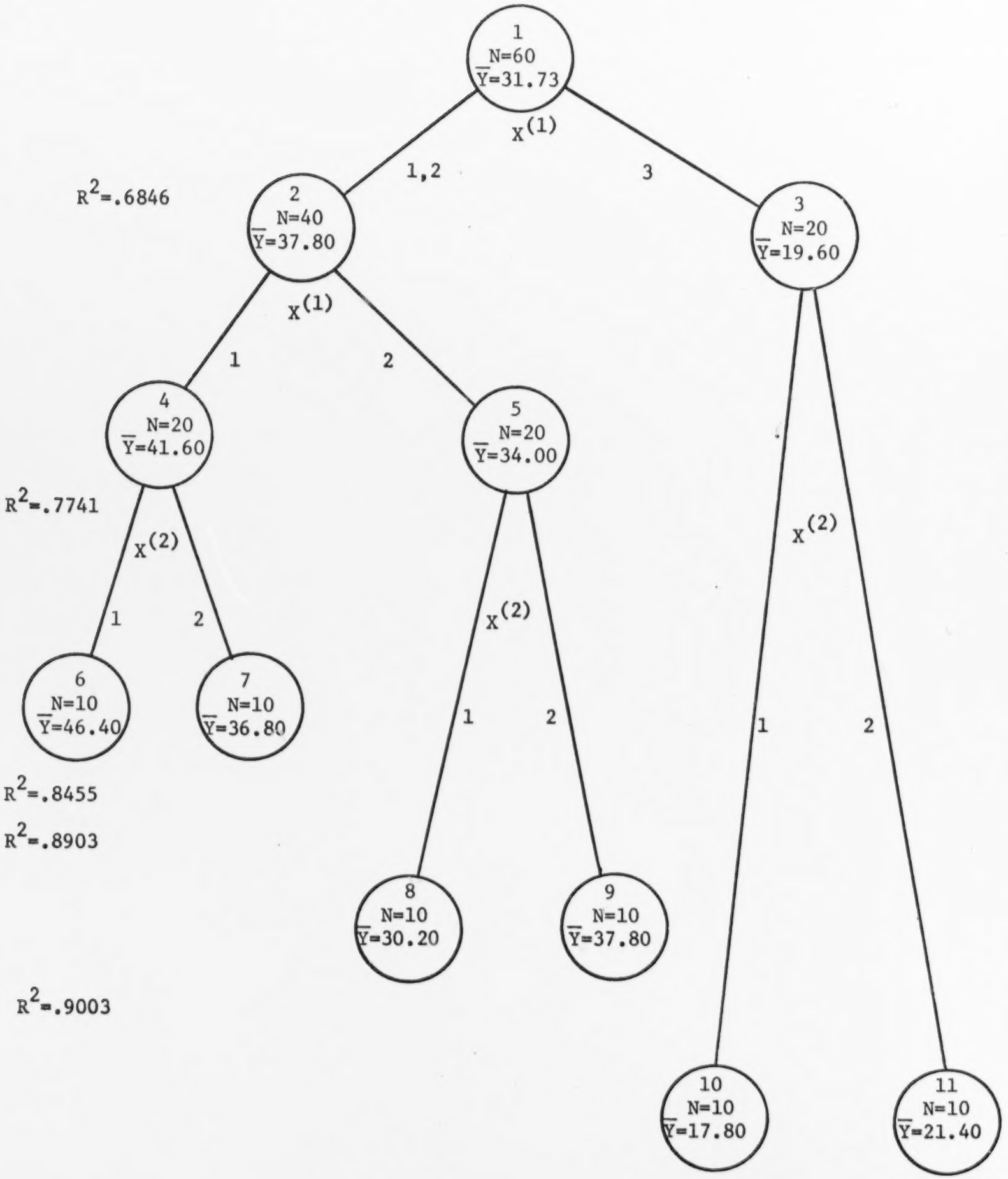
It should be obvious at this point that had we had a more complex problem, for example 40 predictor variables with 10 levels each, the guesswork would have been futile and totally unreasonable. The number of possible mutually exclusive categories in the model would be $10^{40}$, most of which would be empty, considering that the total population of the earth is approximately $4 \times 10^9$.

This was the reason for implementing and developing AFHRL's version of AID-4. The algorithm of AID-4 is a reversal of the model building process. Rather than starting with a full model, including all possible predictors and their simple and complex interactions, AID-4 starts with the ultimate restricted model, namely, the whole group as a unit. By a unique splitting process maximizing the between sum of squares (BSS) for the categories of each variable while minimizing the error sum of squares (within group sum of squares) AID-4 seeks out that variable which has the largest BSS and splits the original group into two mutually exclusive groups on this variable at that category where the maximum BSS occurred. For example, given an 80 variable problem with 10 categories per variable, if the maximum BSS was found in Variable 9 and between categories 1, 2, 3 and 4, 5, 6, 7, 8, 9, 10; the original Group 1 will be split into two

mutually exclusive groups: (a) Group 2 consisting of those individuals whose response to Variable 9 was 1 or 2 or 3, and (b) Group 3 consisting of the remainder of the individuals whose response to Variable 9 was 4, 5, 6, 7, 8, 9 or 10. In actuality, AID-4 has identified the first level full model consisting of 2 groups. The test of significance is an F-test comparing the minimized error sum of squares of the full model (2 groups) and the restricted model (original 1 group). The test of significance for the first split is equivalent to an F-test obtained by a one-way analysis of variance comparing the 2 groups on the criterion variable. The process continues until a specified stop-criterion is reached. Each time a split occurs, the resulting j mutually exclusive groups represent the full model, and the minimized error sum of squares of this model is compared with the error sum of squares of the previous model, consisting of (j-1) mutually exclusive groups. The final split represents an optimal full model which could have been hypothesized before starting to impose restrictions. Going from the final model with the last split towards the original unsplit group, each unsplit group represents an additional restriction.

For our example, the AID-4 splitting process is illustrated in Figure 1. Going down the branches of the tree-pattern, one can identify the simple and complex interactions of the optimum polynomial multiple regression equation. We know that we have predictor variables $x^{(1)}$ and $x^{(2)}$. The first two splits occurred on $x^{(1)}$, $x^{(1)}$ respectively, hence we have an $\left[x^{(1)}\right]^2$ term. The first three splits occurred on $x^{(1)}$, $x^{(1)}$

FIGURE 1. SPLIT DIAGRAM



$x^{(2)}$ respectively, hence we have an $\left[x^{(1)}\right]^2 \cdot x^{(2)}$ term. The second branch from the left is identical to the first identifying the same $\left[x^{(1)}\right]^2 \cdot x^{(2)}$ term. The third branch from the left split on $x^{(1)}$, $x^{(2)}$ respectively, hence we have an $\left[x^{(1)} \cdot x^{(2)}\right]$ term.

Thus, the optimal model is:

Model 6

$$y = p_0 u + p_1 x^{(1)} + p_2 x^{(2)} + p_3 x^{(1)} x^{(2)} + p_4 \left[x^{(1)}\right]^2 + p_5 \left[x^{(1)}\right]^2 \cdot x^{(2)} + e_6$$

which yields, after conventional solution, an $R^2 = .9003$ which is the same as AID-4 arrived at after the final split. Note that Model 6 does not contain a term $\left[x^{(2)}\right]^2$ which is consistent with the previous findings namely that Model 3 and Model 5 were identical (the only difference being that Model 5 contained $\left[x^{(2)}\right]^2$ ).

The major advantage accruing to the task scientist using AID-4 is obtaining the maximum squared composite correlation without the task of attempting to identify the various relevant combinations of linear and non-linear interaction terms by trial and error necessary in the full model of the multiple regression technique. AID-4 automatically identifies these terms. The means of the final categorical groups are the proper weights to be assigned for each of those groups in predicting the criterion variable. An additional major advantage is that out of a regression analysis with a large number of predictor variables, there may be only a small subset of predictor variables which are of significance in the prediction system. AID-4 identifies such a subset of predictors automatically. Finally, the branching pattern facilitates interpretation of the results. In our sample example, it is much more meaningful to

identify Group 6 on Figure 1 as pilots who have advanced academic degrees and who have a predicted score of 46.40, than in a polynomial regression equation where one would have to square "educational level" and multiply it by "pilotness" in order to identify the term $\left[x^{(1)}\right]^2 \cdot x^{(2)}$. In a large prediction system, attempts to identify and include all possible combinations of interaction terms represents a practical impossibility without the help of AID-4.

Many additional and useful bits of information are provided by the output of AID-4, some of which are: (1) at each split, the increased present total explained variance ($R^2$) is printed, together with a statistical test of significance for the difference between the error sum of squares of the new model and the previous model prior to the split; (2) the splits occur in a descending order of importance, that is, the first split identifies that variable which contributes the most to the explained variance; the second split identifies the second variable or a subset of the first split as the next most important contributor to the explained variance; and so on. This hierarchy is very helpful especially if after a few splits a reasonably high $R^2$ is obtained, thus giving the researcher an option of using only a few predictors in the prediction system; (3) the branching pattern of splits reflects trends of characteristics specific to the groups split; that is, it can serve as an "eyeball" pattern analysis. Following the path of each branch of the split-tree, one can identify major characteristics of the final groups on which they differ the most in light of the criterion measure;

(4) cross-validation and double cross-validation options which either splits the original sample into two random samples or takes two given samples, treats each sample separately, determining an optimal split pattern for each and the associated $R^2$. Then it forces the split pattern of Sample 1 upon Sample 2 and vice-versa computing a squared composite correlation for these forced splits. The differences between the optimal $R^2$ for each sample and the corresponding squared composite correlation obtained by forced splitting is a good indicator of the stability of the system; (5) selective or "partial" effects of the predictors are identified such that even if the so-called "main effect" of a particular variable in a complex analysis of variance results in a non-significant F-ratio, AID-4 selectively indicates the level on the other variable(s) at which this non-significant effect becomes significant.

Copies of the write-up and program (to be loaded on a tape provided by the user) can be obtained by written request from Dr. Janos Koplyay, Chief, Statistical and Computer Technology Section, AFHRL/PHSM, Lackland AFB, Texas 78236.

## References

Bottenberg, R. A., and Ward, J. H. Applied multiple linear regression. PRL-TDR-63-6. Lackland AFB, Tex.: 6570th Personnel Research Laboratory, Aerospace Medical Division, March 1963.